How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets

Chung Peng Lee[1], **Rachel Hong**[2], **Harry Jiang**[3], **Aster Plotnik**[4], **William Agnew**[3], **Jamie Morgenstern**[2]

[1]Princeton University [2]University of Washington [3]Carnegie Mellon University
[4]University of Toronto

cl6486@princeton.edu

Abstract

The internet has become the main source of data to train modern text-to-image or vision-language models, yet it is increasingly unclear whether web-scale data collection practices for training AI systems adequately respect data owners' wishes. Ignoring the owner's indication of consent around data usage not only raises ethical concerns but also has recently been elevated into lawsuits around copyright infringement cases. In this work, we aim to reveal information about data owners' consent to AI scraping and training, and study how it's expressed in DataComp, a popular dataset of 12.8 billion text-image pairs. We examine both the *sample-level* information, including the copyright notice, watermarking, and metadata, and the *web-domain-level* information, such as a site's Terms of Service (ToS) and Robots Exclusion Protocol. We estimate at least 122M of samples exhibit some indication of copyright notice in CommonPool, and find that 60% of the samples in the top 50 domains come from websites with ToS that prohibit scraping. Furthermore, we estimate 9-13% with 95% confidence interval of samples from CommonPool to contain watermarks, where existing watermark detection methods fail to capture them in high fidelity. Our holistic methods and findings show that data owners rely on various channels to convey data consent, of which current AI data collection pipelines do not entirely respect. These findings highlight the limitations of the current dataset curation/release practice and the need for a unified data consent framework taking AI purposes into consideration. Code is available at https://github.com/Anderson-Lee-Git/tracing-data-consent-in-datacomp.

1 Introduction

Web-scraped vision-language datasets (VLD) comprising billions of samples have enabled the success of CLIP [1] as well as text-to-image models like Stable Diffusion v1 [2], DALL-E [3], and MidJourney [4]. However, the reliance on copyrighted material from the web to train foundation text-to-image or vision language models remains the subject of much recent debate, especially in recent lawsuits against OpenAI, Stability AI, and Meta[1]. While efforts toward transparent use of copyrighted training data have been explored in text-based pre-training datasets [5, 6], the data consent landscape of web-scraped VLDs remains relatively underexplored, especially as multimodal image-text models become increasingly common.

[1]*Andersen v. Stability AI*, No. 3:23-cv-00201 (N.D. Cal.), *Getty v. Stability AI* [2025] EWHC 38 (Ch), *Kadrey v. Meta*, Nos. 3:23-cv-03417, 3:24-cv-06893 (N.D. Cal.), *NYT v. Microsoft*, No. 1:23-cv-11195 (S.D.N.Y.)

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The shift from the text modality to the image-text modality results in several changes in data consent mechanisms: (1) The signals of data consent in image-text samples are heterogeneous, and (2) image content is often delivered via third-party cloud providers, making the practice of tracking data provenance more challenging. Despite these changes, the impact of violating data consent in the vision-language landscape is no less concerning than that in the text-based counterpart, especially as visual artist communities have spoken out about potential economic loss and reputational harm as a result of generative AI systems [7].

Furthermore, in recent cases involving Anthropic and Meta[2], although the training on copyrighted material was deemed "fair use," the alleged collection of content from pirated sources remains contentious and has precluded the dismissal of the case. This decision raises questions around how dataset curation methods gather data in the first place, and whether such sourcing is allowed. In light of the lack of transparency in web-scraped VLD's data consent [8], we aim to *demystify the data consent mechanisms throughout the life cycle of curating, releasing, and using a web-scraped VLD.*

Specifically, we use DataComp's CommonPool [9] as a case study of the web-scraped VLDs. They sourced image-text pairs from CommonCrawl [10], an archive of web pages crawled from the internet, and performed deduplication and minimal filtering to produce a set of 12.8B *url-text* pairs, where the *url* points to the image content. As of July 2025, CommonPool has over 2M downloads [11]. Pulling from the same web archive, CommonPool has substantial overlap with its precursor, LAION-5B [12], which enabled the early version of Stable Diffusion v1, MidJourney, and Google's Imagen [2, 4, 13]. Even though the data used to train OpenAI's CLIP or DALL-E were not disclosed, the corresponding papers claim to have sourced the training datasets from the internet [1, 3], similar to CommonPool. Therefore, we believe CommonPool as a case study not only informs the open-source vision-language model development community but also provides a lens into commercially protected datasets.

We recognize and take advantage of various signals provided by the image, text, metadata, and their associated data host. We use both sample-level characteristics, such as copyright notice, the exchangeable image file format (EXIF) [3] metadata, and watermark detection, and web-domain-level characteristics, such as Terms of Service (ToS) and Robots Exclusion Protocols (REP), also known as robots.txt. We make the following contributions:

1. Investigate data consent mechanisms in a web-scraped VLD provided by the information in the released artifact
2. Estimate approximately 122M of samples in CommonPool have included copyright information, and over 60% of samples from the top 50 domains, in the `small-en` scale of CommonPool, are sourced from sites restricting scraping in their ToS.
3. Demonstrate that data owners often rely on inconsistent channels to convey data consent, of which AI data collection pipelines do not fully respect, surfacing issues of a lack of a uniform consent mechanism.
4. Use our findings to outline various limitations and recommendations for future web-scraped VLD curation.

2 Background

2.1 Terminology

Legal discussion around training with web-scraped data involves specific terms; in this section, we outline the scope of each term and the role they play in the explicit permission granted to use the data. We limit our focus to examining data consent and copyright implications within the United States.

Copyright. As defined by the U.S. Copyright Office [14], copyright protects the expression of original work. As long as the work is *fixed*, *expressed in tangible forms*, and not an idea, concept, fact, or other exception, it automatically becomes copyright-protected. Notably, the role of the *copyright notice*, like "© John Doe 2025", is to publicly claim that the work is protected by copyright. As such,

[2]*Kadrey v. Meta* (see *supra.*), Doc. 598 (Partial Summary Judgment), and *Bartz v. Anthropic PBC*, 3:24-cv-05417, (N.D. Cal.), Doc. 231 (Partial Summary Judgment)

[3]https://en.wikipedia.org/wiki/Exif

Table 1: Summary of quantitative results in our measurement of data consent mechanisms, including Copyright Notice, ToS, and Robots.txt.

Mechanism	Finding
Copyright Notice	1. We estimate 122M English-captioned samples in CommonPool to contain copyright information. 2. We estimate the watermark prevalence to be 9–13% with 95% confidence interval.
Terms of Service	1. 33% of samples from top 50 web domains are restricted to personal/non-commercial/research. 2. 60% of samples from top 50 web domains are against scraping.
Robots.txt	1. AI-purposed bots are mostly disallowed in existing robots.txt. 2. We find 28% of samples with observed robots.txt to disallow CommonCrawl crawling (via CCBot), the upstream dataset of CommonPool.

it becomes more difficult for defendants in infringement cases to argue they were not aware of the work being copyrighted [15].

License. A license, or agreement, grants specified rights to someone to use the work for purposes protected by copyright, such as reproduction, display, or making derivatives. A license could be useful for the creator to limit the use of the work in certain scenarios without placing it in the *public domain*, which is outside the scope of copyright protection.

Data Consent. We refer to data consent as the "permission" granted for the user to use the data for model training purposes. This is not limited to any form of written consent, such as ToS, copyright notice, claims, or license. In other words, data consent is obtained when the user follows the acceptable pipeline to retrieve data proposed by the data host or data owner. As an example, even if the data is not copyright-registered through the U.S. Copyright Office, a written ToS to restrict the use of such data for model training purposes would be considered a "restriction to use" in the scope of data consent we consider.

2.2 Involved Parties

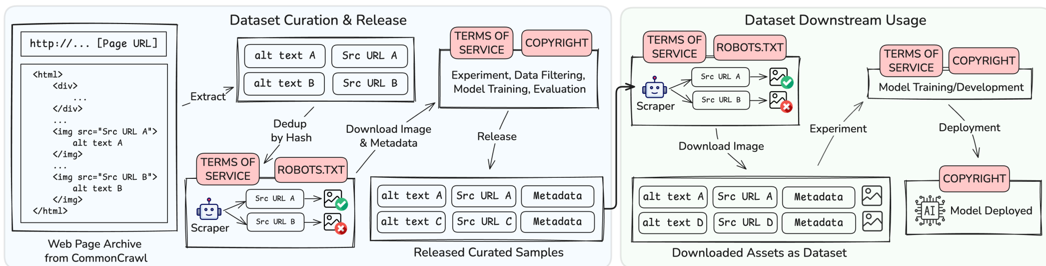
The pipeline to curate, release, and download a web-scraped dataset involves multiple entities. To study the data consent landscape, we first define how the stakeholders are involved in the life cycle of such datasets.

- *Dataset Curator* – The curator of the dataset releases a set of *url-text* pairs for downstream use. In the case of DataComp [9], it would be their authors.
- *Dataset User* – The user of the dataset downloads the pairs of URLs and texts released by the *Dataset Curator*.
- *Data Owner* – The owner of the image data itself. Since tracing data ownership on the internet is extremely difficult, we relax the ownership to be the action of embedding the image on their web page. This relaxation builds on the assumption that the actor of embedding the image respects the copyright of the image and shares it per the level of consent they obtain.
- *Data Host* – The data host is the entity that owns the image URL referred to by the sample. Since the delivery of image content is often optimized through content delivery network (CDN) and cloud providers, this entity may exhibit little information about the *Data Owner*.

2.3 Life cycle of web-scraped VLD

Curation & release. The top-level raw source of data originates from CommonCrawl [10]. The collection of *url-text* pairs comes from extracting the `<img src=URL>alt text</img>` from the internet. This extraction *does not* consider the *page url* where the image appears. Figure 1 illustrates the distinction between *page url* and *src url*. With the extracted *url-text* pairs, the *Dataset Curator*

Figure 1: The life cycle of curating, releasing, and using the web-scraped VLD. Even though the *Dataset Curator* initially downloads the image assets in their curation process, the released samples only contain the caption, *src url* pointing to the image asset, and image metadata. To access the dataset, the *Dataset User* must download the images following the released URLs. The red tags on each step indicate the data consent mechanism we consider involved. The red tag indicates the relevant data consent signals in each component.



uses tools like img2dataset [16] to automatically download all the images from these URLs, referred to as *scraping*. Since the URLs are extracted from archives of the internet, *not all download attempts are successful or align with the original image*. For instance, the owner of the URL could replace the image with another image or take down the image completely. With the downloaded assets, the *Dataset Curator* experiment with data filtering, cleaning, augmentation, and model training/evaluation to curate the best set for release. Finally, the release of the curated dataset comprises *url-text* pairs along with metadata they obtain from either their experiments or downloading, *without the actual image assets.*

Downstream usage. The *Dataset User* first obtains the index of *url-text* pairs released by the *Dataset Curator*. Since the released dataset artifact comes without the image assets, the *Dataset User* has to utilize similar tools to *scrape* through the provided URLs. In the case of DataComp [9], the scraping functionality is provided as part of the release. This mechanism inherits the same drawback of potentially inconsistent or failed downloads. Not only does it potentially diverge from the *Dataset User*'s expectation of the released dataset, but it might also expose the *Dataset User* to the risk of data poisoning [17]. Furthermore, since the *Dataset User* is scraping the web with the index of the URLs, the *Dataset User* is responsible for abiding by any ToS or other data consent mechanism specified by the website hosting the content. With the image assets downloaded, the *Dataset User* then experiments with the downloaded samples in their storage.

3 Methods

We first outline the concrete experiment setup for our audit, including data filtering, sizes, and scales that we audit. Then, we present the methods in two categories, one at the sample level and the other at the web domain level. These two angles allow us to audit how image owners and website owners disclose consent for scraping and AI training.

3.1 Setup

CommonPool was released at four scales: `xlarge` (12.8B), `large` (1.28B), `medium` (128M), and `small` (12.8M), where the largest contains 12.8B image-text pairs and the lower scale is a subset of the larger ones. Due to limited storage space and compute resources, we study both `small` and `medium` such that we can verify whether results found in `small` are also observed in `medium`.

Moreover, since legal mechanisms of data consent are dependent on specific jurisdictions, we restrict our target data to be English-based. Particularly, we follow the same measure in Gadre et al. [9] to use `fasttext` [18] to filter the original dataset by English-only captions. Table 2 summarizes the audited dataset.

Table 2: Sample counts of CommonPool's configurations considered in our work. `scale-en` refers to the English-filtered version of the original scale. Accessible counts refer to images downloadable through the released link. "Top 50" refers to the subset in the top 50 *base domains*.

Scale	Released	Accessible	"Top 50"
`small`	12.8M	9.8M	–
`small-en`	6.3M	4.8M	2.1M
`medium`	128.0M	98.3M	–
`medium-en`	63.0M	47.7M	21.5M

Table 3: Number of samples found through each measurement method, where Caption and OCR refer to searching the copyright notice through samples' captions and OCR-extracted texts.

Measure	`small-en`	`medium-en`
Caption	10,585 (0.22%)	98,555 (0.21%)
OCR	4,307 (0.09%)	38,697 (0.08%)
EXIF Metadata	108,951 (2.27%)	1.09M (2.28%)
Caption $\cup$ OCR $\cup$ EXIF	123,096 (2.56%)	1.22M (2.55%)

3.2 Sample-level Characteristics

At the sample level, we use text, visual, and metadata information to source characteristics of data consent. Particularly, we search for samples with the presence of *copyright notice*, *copyright field in metadata*, and *image watermark*. With the presence of this information, it becomes difficult for a defendant on copyright infringement to argue ignorance of the fact that the material was copyright-protected [15].

Copyright Notice. We crafted a set of regular expressions to capture common copyright notices such as "©" and "copr." These rules are applied to both caption and OCR-extracted text, where we use open-source PaddleOCR [19] for extraction. The full list of search patterns is included in Appendix Section B.

Copyright Field in Metadata. *Exchangeable image file format* (EXIF) is a standard of image metadata to specify information about the image itself as well as the digital device that produced the image. For instance, some tags include original height, width, focal length, and color space. We search for samples of which the metadata contains a non-empty copyright tag field keyed by "Copyright" or "0x8298," following the EXIF standard version 2.3. [20].

Image Watermark. A watermark detection classifier aims to output whether or not a given image contains a watermark. We (1) use off-the-shelf watermark-finetuned YoloV8 [21, 22], (2) build a watermark-finetuned MobileViTv2 [23], (3) use two SOTA open-source VLMs, Rolm OCR [24] and Gemma-3-12b-it [25] as our detection methods. To validate the faithfulness of these methods, we evaluate them on (1) *watermark-eval*: Felice Pollano [26]'s validation set, with a balance of $\sim$3200 images for both watermarked and non-watermarked images, and (2) *datacomp-watermark-eval*: a random 955-image subset of CommonPool we annotate, to validate the robustness of our detection methods on web-scraped images. Last but not least, we question the faithfulness of LAION-5B's release of *watermark score*[4] by annotating a subset of LAION-5B and analyzing the utility of those scores. The full training and evaluation details can be found in Appendix Section A.

3.3 Web-domain-level Characteristics

At the web-domain level, the administrator who hosts the content typically specifies rules on permitted usage of their content. Particularly, we examine the top 50 web domains' ToS and their REP, which specifies the restriction of scraping/crawling bots. The top 50 domains are defined by the counts of samples sourced from these domains. In both `small-en` and `medium-en` scales, the top 50 domains cover $\sim$45% of all samples, namely 2.1M and 21.5M samples respectively.

The web domains are extracted from *src url* as provided by CommonPool, which points to the image asset, rather than the original website where the content is embedded, which we call *page url*. Furthermore, since most content is delivered through domains designed for static content or a content delivery network (CDN), we extract the *base domain* by trimming off the prefix to aggregate the sharded domain URLs. For instance, Pinterest uses bucketed web domains like `i.pinimg.com` and `i-h1.pinimg.com` to deliver content. Through extracting only the *base domain*, which would

[4]LAION-5B releases watermark scores per sample to estimate the probability of the presence of watermark in the image.

be `pinimg.com` in the example, we have a more accurate estimate of sample counts for each web domain.

Terms of Service (ToS). Following Longpre et al. [5], we annotate each web domain with the following attributes: (1) Category: the core function of the *Data Host*, (2) License Type: the permission granted to the end user, and (3) Scraping Policy: the restriction on web-scraping. In this work, we focus on the act of *scraping*, the action of automatically downloading/copying a vast majority of data through an index of links, because both the *Dataset User* and *Dataset Curator* directly engage in this act.[5]

Similar to Fiesler et al. [27]'s qualitative analysis process, we have two coders to annotate each web domain's attributes, but we start with the codebook for (2) and (3) from Longpre et al. [5]. For the Category, the primary coder first builds the codebook when iteratively going through the web domains. After creating the initial codebook and first pass, the second coder annotates the web domains. The two coders resolve any conflict through adjusting either the annotations or the codebook. Table 8 in the Appendix summarizes the types in each attribute, and the full codebook is included in Appendix Section C.

Robots Exclusion Protocol (REP). REP, implemented via robots.txt, allows website administrators to specify which automated clients (user agents) can access their sites. Administrators can allow or disallow access for specific agents, such as "CCBot" (CommonCrawl), "GPTBot" (OpenAI), or any agent using the wildcard "*". They can also restrict access to certain website paths. In Germany, robots.txt is legally enforceable, with exceptions for scientific research [28, 29].

For each of the top 50 *base domains*, we map the *base domain* to a list of *full domains*, which are the web domains with the original prefix. For instance, the *base domain*, `pinimg.com`, maps to a list of *full domains*, `[i.pinimg.com, i-h1.pinimg.com, ...]`. We retrieve robots.txt by appending "robots.txt" at the end of the *full domains*. In the `small-en` scale, there are 96,436 unique URLs requested, and 81,273 (84%) of them successfully return with a non-empty robots.txt[6].

We parse each robots.txt following Longpre et al. [5] to three categories: *All Disallowed*, *Some Disallowed*, and *None Disallowed* for agents listed in the robots.txt file. *All Disallowed* is when "a particular agent is mentioned and disallowed from all parts of the site." *None Disallowed* is when "the particular agent is mentioned and allowed for all parts of the site," or "has no disallowed parts." *Some Disallowed* is when "a particular agent is mentioned and disallowed from some parts of the website." An agent must be listed in robots.txt to determine the category, and we consider every agent appearing in any of the parsed robots.txt.

4 Results

In this section, we present our findings according to the sample-level and web-domain-level methods of determining data consent.

4.1 Sample-level Statistics

Approximately 122M English samples contain characteristics of copyright notice or claims in CommonPool.

We find 1.22M samples exhibiting characteristics of copyright notice or claims in the `medium-en` scale. We further validate the faithfulness as the portions of the found samples through each method scale similarly from `small-en` to `medium-en`, as shown in Table 3. This extends our results to implications on the full dataset of 12.8B samples, where approximately 122M of English samples may contain copyright notices or claims. We observe very little overlap between the keyword search methods across image, text, and EXIF metadata. This signifies that copyright claims are heterogeneously disclosed for images on the internet, which emphasizes the need to examine each modality to adequately determine copyright information from web-scraped samples.

[5] In contrast, the term *crawling* refers to the act of developing a spider to recursively follow links from web pages to store content.

[6] In the `medium-en` scale, there are 434,498 URLs requested, and 392,286 of them successfully return with a non-empty robots.txt.

Table 4: Evaluation of watermark detection methods on both standard watermark detection dataset, *wm-eval* with 3289 clean and 3299 watermark images, and an annotated set of web-scraped images from CommonPool, *datacomp-wm-eval* with 849 clean and 106 watermark images.

Model	*wm-eval*			*datacomp-wm-eval*		
	Precision	Recall	F1	Precision	Recall	F1
Finetuned YoloV8	97.44	95.90	96.66	42.63	51.88	46.80
Finetuned MobileViTv2	90.43	86.63	88.49	11.02	74.53	19.20
Rolm-OCR	99.15	49.74	66.25	50.80	59.43	54.78
Gemma-3-12b-it	99.22	81.87	89.71	41.05	73.58	52.70

Watermarks are present in web-scraped images, but detecting them remains a major challenge — even for state-of-the-art methods.

In our evaluation suites, we use (1) *watermark-eval*, comprising a balance of 3289 clean and 3299 watermarked images, and (2) *datacomp-watermark-eval*, a random sample of 955 images from CommonPool we annotate. We find that 106 of *datacomp-watermark-eval*, or 11.09%, are watermarked, resulting in a 9% to 13% of the distribution with 95% confidence interval. From Table 4, we observe that across all models, the F1-score significantly drops on *datacomp-wm-eval*. This indicates a distribution shift between the traditional watermark detection dataset and the web-scraped images "in the wild." Upon investigation, we determine that traditional methods tend to have lower precision on *datacomp-watermark-eval* because of the text appearing in the image, where the models tend to output *True* for images with texts in them.

Is LAION-5B's released watermark score faithful or informative for understanding and respecting data consent?

In light of our watermark detection experiments, we question the fidelity of the watermark score released in LAION-5B [12]. We annotate 1308 random samples from LAION-5B and find that 176 have a watermark, or 13.45%. Furthermore, using the standard threshold of 0.5 on the watermark scores released, the precision and recall are only at 34.09% and 51.13%. The area under the receiver operating characteristic (ROC) curve is 0.74. These statistics further demonstrate the difficulty of watermark detection for web-scraped images "in the wild" observed in our experiments. Moreover, the low performance of LAION-5B's watermark score reveals the low utility of this watermark probability score if a dataset user wishes to avoid training AI systems on watermarked images.

4.2 Web-domain-level Statistics

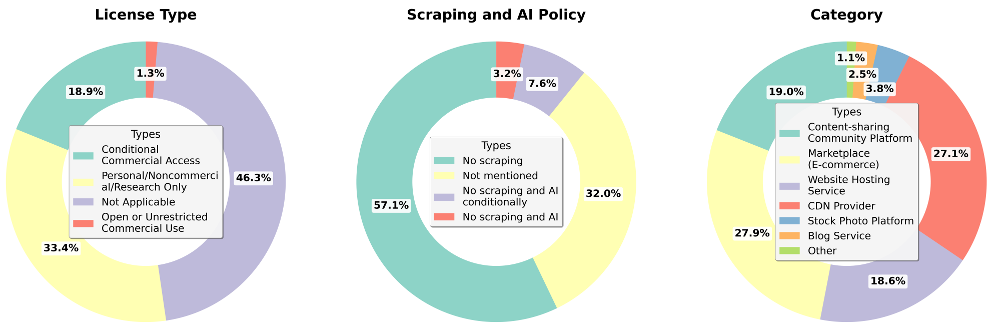
Since the top 50 base domains in `small-en` and `medium-en` only differ by 1 base domain, we present the results for `small-en` for conciseness. The distribution of the top 50 base domains can be found in Figure 4 in the Appendix. For robots.txt, we primarily present our results with the top six user agents in terms of the number of "observations," or samples that come from sites with robots.txt files that mention the top six agents. The total number of observed agents, weighted by sample counts, is 1.1M. Full results are included in Appendix Section D.

60% of samples in the top 50 base domains prohibit scraping, and 33% of them are restricted to Personal/Research/Non-commercial Only Use.

Through our analysis of the ToS in Figure 2, 57.1% of the top 50 base domains prohibit general scraping without mentioning AI, and 3.2% prohibit scraping and AI unconditionally. This not only emphasizes the responsibility of the *Dataset Curator* but also that of the *Dataset User*, who scrapes these sites as well while downloading CommonPool. Furthermore, 33.4% of samples in the top 50 base domains come from websites with ToS limiting usage of content for Personal/Research/Non-commercial purposes.

The practice of releasing only url-text pairs restricts the ability to examine data consent through ToS.

Figure 2: Terms of Service annotations. The full population in each chart is all samples in the top 50 base domains of `small-en`. The portion is determined by the exact number of samples in each type. For License Type, "Not Applicable" indicates that the ToS from the base domain does not specify or provide any license type information. For Category, "Other" indicates that the base domain is for a very domain-specific service. For instance, *4sqi.net* is delivered by Foursquare, a location-intelligence service provider.



Web-scraped VLDs, such as CommonPool, LAION-400M, and LAION-5B, all use the practice of releasing only the *src url* and caption as described in Section 2. We find that 27.1% and 18.6% of the samples in the top 50 base domains are under CDN Provider and Website Hosting Service categories, respectively. Yet, the ToS of `amazonaws.com` cannot fully reflect the actual ToS used by the website offering the content stored at those *src urls*. The core reason is that image content delivered via *src url* is often through a CDN or static content host, and only those *src urls* are released instead of the original *page url*. Without the context of *page url*, the website URL where the *url-text* pair is extracted, a thorough examination of data consent is infeasible. This characteristic also primarily accounts for the reason why 46.9% of samples' License Type in Figure 2 are categorized as "Not Applicable," meaning that the provided *src urls*' base domain's ToS may not have the right to specify the License Type.

robots.txt is mostly adopted to convey restrictions for AI-purpose scrapers/crawlers.

In the top 6 agents by number of samples covered by observations, we see that traditional web-indexing (googlebot-image) or wildcard ($*$) agents don't have very high *All Disallowed* rate compared to agents related to AI-purposes such as GPTBot, Bytespider, and claudebot. This phenomenon implies that the website administrator disallowing these AI-purpose agents wishes to prevent the use of their content for model development. However, a dataset user downloading CommonPool to train a model does not specify the user agent by default and therefore can bypass REP to scrape many of these same samples from sites that ban GPTBot, Bytespider, and claudebot. Only 3.9% of samples come from sites that disallow any agent, so many sites that specifically block AI-purpose bots may miss dataset users scraping open-source VLDs to train models.

Moreover, even though CommonPool is sourced from CommonCrawl, which respects robots.txt when sourcing the web pages, we still observe CCBot in 353K robots.txt. The most likely reason is that the user adopts robots.txt to revoke their consent after CommonCrawl archives their pages. Despite this adoption, the collection of CommonPool as an index of url-text pairs continues to direct scraping traffic to those websites that chose to revoke consent when the *Dataset User* downloads CommonPool using a non-CCBot user agent name.

5 Discussion

5.1 Limitation of Current Release Practice

Problem. Our results reveal several drawbacks in the current release practice of web-scraped VLDs. Firstly, the lack of *page url* greatly restricts the ability to probe whether an image is prohibited from use by the associated ToS. This issue originates from a combination of how image content is usually delivered through CDN, how each sample is collected by only an HTML tag, and how the website

Table 5: Top results from robots.txt analysis for `small-en` scale's top 50 *base domains*, accounting for 96,436 attempted *full domains*, 81,273 successful robots.txt, and 1,126,876 samples observed. For each agent, the number of observed cases is broken down by the number and percentage (relative to observed) of cases where all, some, or none were disallowed. The dark gray background highlights rows that have over 80% *All Disallowed* rate, and the 🤖 icon indicates that the agent is AI-purposed. "All Agents" row refers to an aggregation of all agents found in all the examined robots.txt. The aggregation rule is as follows: If for all agents, a robots.txt has *All Disallowed*, then the decision is *All Disallowed*. If for any agent in all agents, a robots.txt has *All Disallowed* or *Some Disallowed*, then a robots.txt has *Some Disallowed*. Otherwise, it has *None Disallowed*.

Agent	Observed	*All Disallowed*		*Some Disallowed*		*None Disallowed*	
		Count	% of observed	Count	% of observed	Count	% of observed
"All Agents"	1,126,876	6,442	0.6%	1,014,576	90.0%	105,858	9.4%
GPTBot 🤖	578,498	538,431	93.1%	40,028	6.9%	39	0.0%
*	475,139	18,595	3.9%	391,799	82.5%	64,745	13.6%
CCBot 🤖	353,324	313,920	88.8%	39,365	11.1%	39	0.0%
Bytespider 🤖	301,344	262,029	87.0%	39,274	13.0%	41	0.0%
googlebot-image	224,268	0	0.0%	224,166	100.0%	102	0.0%
claudebot 🤖	224,200	224,199	100.0%	1	0.0%	0	0.0%

itself (*page url*) is not always related to the extracted HTML tag. Secondly, releasing an index of the web through *url-text* pairs allows the *Dataset Curator* to avoid hosting any image asset, and thus any copyright infringement claim or responsibility of providing a convenient channel for the *Dataset User* to access the copyrighted/restricted-to-use data. This shift of accountability may not be made aware to the *Dataset User*, creating an illusion that the curation of an open-sourced web-scraped VLD has already dealt with data consent, so usage of that dataset is in the clear.

Recommendation. For better data provenance and transparency, we recommend that future releases include the website page where the samples are collected. Moreover, the *Dataset Curator* should either *clearly inform or warn* the *Dataset User* about the potential responsibility of scraping when using their dataset, or take the responsibility to construct the dataset with standalone image assets respecting the *Data Owner*'s consent, through the various mechanisms we used in our audit.

5.2 Call for a Unified Data Consent Framework

Problem. In our case study of DataComp CommonPool, we find that each audit approach surfaced a distinct set of samples restricting data usage with very few overlaps. This observation indicates that the data consent is conveyed through multiple channels, such as image metadata, copyright notice, or image watermark. Even though this highlights the importance of auditing through our comprehensive techniques, it presents a problem of lacking a universally recognized framework to convey data consent, particularly in the life cycle of AI data collection. For instance, robots.txt was constructed for web scraping, but web scraping is only a part of the life cycle. As another example, the copyright notice goes beyond the consent for model development, but also for display, re-distribution, and so on. In addition to the divergent channels to convey data consent, Longpre et al. [5] reveals a contradiction between these channels where ToS have different restrictions from REP.

Recommendation. All the involved parties highlighted in this work need a common protocol such that data owners can communicate data consent, specifically for the use of model development. The Robots Exclusion Protocol is not sufficient because we showed that website maintainers often are not the owners of the data. We believe that a unified channel not only helps the *Data Owner* to protect their works from misuse, but also guides the *Dataset Curator* and *Dataset User* to respect their data consent. Such a framework should not only be adopted but also treated as the source of truth to represent data consent. In addition, we encourage the adoption of an opt-in understanding of consent, as supported by many data owner stakeholders [30, 31]. Proposed solutions such as Spawning [32], an opt-out model, do not address the obscurity of scraping and training to many data owners, and implicitly obfuscate consent.

6 Related Work

Prior work on auditing web-scale pre-training datasets ranges from data governance, privacy to social biases encoded. In the text modality, Dodge et al. [33] highlighted the importance of documenting these datasets with the excluded data's characteristic, web domain distribution, and other aspects of Colossal Clean Crawled Corpus (C4) [34]. Elazar et al. [6] extended the goal to understand these datasets to several pre-training datasets, such as C4, LAION-2B-en, and The Pile [34, 12, 35]. by documenting their domain statistics, contamination with evaluation sets, and PII inclusion. More specific to data consent, Longpre et al. [5] investigated the consent mechanism of text-based pre-training datasets including C4, dolma, and RefinedWeb [34, 36, 37]. They focus on the temporal changes in data consent in both ToS and robots.txt and highlight the increasing restrictions on the web to train AI models with web-scraped data.

In the vision-language datasets landscape, Hong et al. [38] studied the impact of data filtering on the exclusion/inclusion statistics concerning minority groups across gender, religion, and race. Hong et al. [39] presented a legally-grounded study on private information existing in CommonPool and its implications from a legal perspective. Our work studies the data consent mechanism in the landscape of web-scraped VLDs.

7 Acknowledgement

We would like to thank Christina Yeung for the thoughtful feedback on licensing, policy, and the writing revisions. This research is supported by the NSF Graduate Research Fellowship Program.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[4] Midjourney. Midjourney. `https://www.midjourney.com/home`, 2025. Accessed: 2025-07-04.

[5] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, et al. Consent in crisis: The rapid decline of the ai data commons. *Advances in Neural Information Processing Systems*, 37:108042–108087, 2024.

[6] Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=RvfPnOkPV4`.

[7] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 363–374, 2023.

[8] Jack Hardinges, Elena Simperl, and Nigel Shadbolt. We must fix the lack of transparency around the data used to train foundation models. *Harvard Data Science Review (Special Issue 5). https://doi. org/10.1162/99608f92. a50ec6e6*, 2024.

[9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

[10] CommonCrawl. Commoncrawl. `https://commoncrawl.org`, 2025. Accessed: 2025-07-04.

[11] Huggingface. Hugginface api. `https://huggingface.co/api/datasets/mlfoundations/datacomp_pools?expand%5B%5D=downloads&expand%5B%5D=downloadsAllTime`, 2025. Accessed: 2025-07-04.

[12] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[14] U.S. Copyright Office. What is copyright. `https://www.copyright.gov/what-is-copyright/`, 2025. Accessed: 2025-07-07.

[15] U.S. Copyright Office. Copyright notice. `https://www.copyright.gov/circs/circ03.pdf`, 2021. Accessed 2025-07-27.

[16] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. `https://github.com/rom1504/img2dataset`, 2021.

[17] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 407–425. IEEE, 2024.

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[19] PaddlePaddle Authors. Paddleocr, awesome multilingual ocr toolkits based on paddlepaddle. `https://github.com/PaddlePaddle/PaddleOCR`, 2020.

[20] Standardization Committee. Exchangeable image file format for digital still camera: Exif version 2.3. `https://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf`, 2012. Accessed: 2025-07-04.

[21] ultralytics community. ultralytics. `https://github.com/ultralytics/ultralytics`, 2025. Accessed: 2025-07-04.

[22] mnemic. mnemic/watermarks_yolov8. `https://huggingface.co/mnemic/watermarks_yolov8`, 2024. Accessed: 2025-07-04.

[23] Sachin Mehta and Mohammad Rastegari. Separable self-attention for mobile vision transformers. *arXiv preprint arXiv:2206.02680*, 2022.

[24] Reducto AI. Rolmocr: A faster, lighter open source ocr model, 2025.

[25] Gemma Team. Gemma 3. 2025. URL `https://goo.gle/Gemma3Report`.

[26] Felice Pollano. Watermarked / Not watermarked images. https://www.kaggle.com/datasets/felicepollano/watermarked-not-watermarked-images/data, 2019. A suite of images with and without a random watermark, divided into training and validation sets. Dataset licensed under CC BY-NC-SA 4.0. Accessed: 2025-07-08.

[27] Casey Fiesler, Cliff Lampe, and Amy S Bruckman. Reality and perception of copyright terms of service for online content creation. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1450–1461, 2016.

[28] L. G. Hamburg. Urteil vom 27.09.2024 - 310 o 227/23. https://openjur.de/u/2495651.html, September 2024.

[29] Official Journal of the European Union. Directive (eu) 2019/790 of the european parliament and of the council of 17 april 2019 on copyright and related rights in the digital single market and amending directives 96/9/ec and 2001/29/ec. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019L0790, 2019. Accessed 2025-07-28.

[30] Lin Kyi, Amruta Mahuli, M. Six Silberman, Reuben Binns, Jun Zhao, and Asia J. Biega. Governance of generative ai in creative work: Consent, credit, compensation, and beyond. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3713799. URL https://doi.org/10.1145/3706598.3713799.

[31] Cultural Intellectual Property Rights Initiative. Consent Credit Compensation: The Legal Literacy Campaign, 2017. URL https://www.culturalintellectualproperty.com/the-3cs.

[32] Spawning. Spawning. https://spawning.ai, 2025. Accessed: 2025-07-31.

[33] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[35] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[36] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, 2024.

[37] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=kM5eGcdCzq.

[38] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *EAAMO*, 2024.

[39] Rachel Hong, Jevan Hutson, William Agnew, Imaad Huda, Tadayoshi Kohno, and Jamie Morgenstern. A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset. *arXiv preprint arXiv:2506.17185*, 2025.

[40] MFW. W6_janf dataset. `https://universe.roboflow.com/mfw-feoki/w6_janf`, jan 2024. URL `https://universe.roboflow.com/mfw-feoki/w6_janf`. visited on 2025-07-09.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

A Watermark Detection Details

A.1 Models

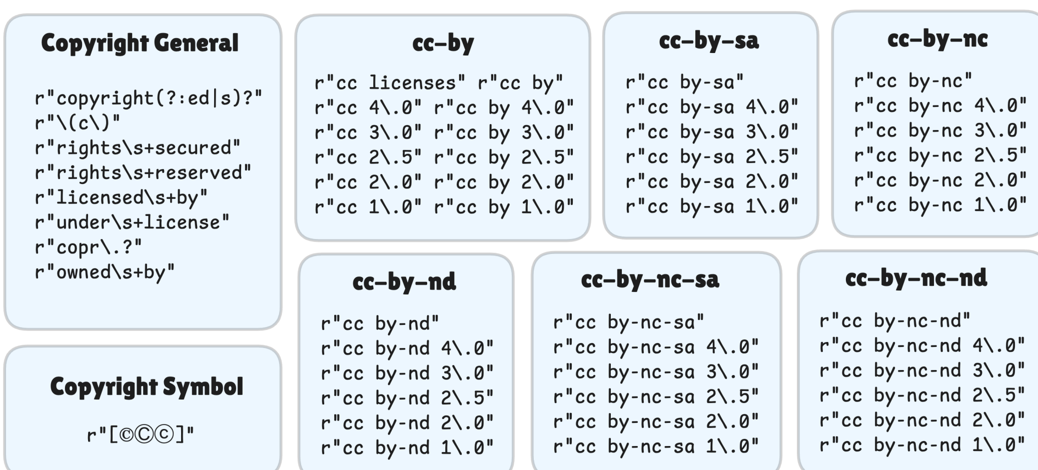
The off-the-shelf YoloV8 is finetuned on MFW [40] comprising 4,935 watermarked images by mnemic [22]. We finetune a pre-trained MobileViTv2 [23] on the training split of Felice Pollano [26] comprising 12,510 and 12,477 images for watermarked and non-watermarked images. The pre-trained MobileViTv2 [23] is loaded via Huggingface checkpoint `apple/mobilevitv2-1.0-imagenet1k-256`. The finetuning is done with a learning rate of 1e-4, weight decay of 0.01, and optimized through AdamW [41]. We use Huggingface checkpoints for both Rolm-OCR and Gemma-3-12b-it, and we prompt the VLMs with: *A watermark on an image is a deliberately embedded visual marker — often semi-transparent text, logos, or patterns — designed to assert ownership, deter unauthorized use, or signal authenticity. It can also be a form of a link, brand name, or author name at the top/bottom corner of the image. Does this image contain any watermark? If so, return the text of the watermark. Otherwise, return no in lowercase.*

A.2 Compute Resources

All model training and evaluation use 2 Nvidia A100 GPUs. The evaluation of VLMs can be done in 40 minutes with one GPU. For finetuning MobileViTv2, the training for 5 epochs can be done in 2 hours with one GPU.

B Copyright Notice Search Pattern

Figure 3: Regular expression search patterns used to source copyright notice in samples' captions and OCR-extracted texts.



The full copyright notice search patterns are illustrated in Figure 3. Each category has multiple regular expression patterns. We find samples that have at least one match for any regular expression in the list. For "Copyright General," we include commonly used patterns to claim copyright. For "Copyright Symbol," we include three encoding variants of copyright symbols for better capture. For the Creative Commons, we search for all 6 license types under Creative Commons, including the past versions.

C Terms of Service Analysis Codebook

There are three attributes we annotate for each web domain: (1) Category, (2) License, and (3) Scraping Policy. Table 8 summarizes the types included in each attribute. The codebook finalized for each attribute and type is as follows:

1. *Category*

- **Marketplace (E-commerce)** – Platforms where *general* goods or services are bought and sold.
- **CDN Provider** – Content Delivery Network *providers* and *services* that deliver web content to users based on geographic location. For instance, `alicdn` and `cloudfront.net` fall under this type. This type does not include CDN incorporated by specific and mappable entities for faster content delivery. For instance, Adobe has its own CDN web domain to deliver its content instead of serving others' content.
- **Website Hosting Service** – Services providing infrastructure for websites to be hosted and accessible on the internet. For instance, `wixstatic.com` and `wp.com` fall under this type.
- **Blog Service** – Platforms for users to publish blogs. For instance, `blogspot.com` falls under this category.
- **Stock Photo Platform** – Platforms where *image assets* are bought and sold, typically under licensing agreements. This type differs from **Marketplace (E-commerce)** in that the *goods* are *image assets* themselves.
- **Content-sharing Community Platform** – Platforms for exchanging user-generated and community-purposed content, as opposed to transaction-based exchange.
- **Other** – Uncommon websites or services that don't fall under any previous category. For instance, `4sqi.com` offers location-intelligence information through its API.

2. *License Type*
   - **Personal/Noncommercial/Research Only** – Use of content is limited to personal, research, or noncommercial contexts. Commercial use is explicitly prohibited.
   - **Conditional Commercial Access** – Commercial use is permitted under certain conditions, such as requiring permission, excluding third-party redistribution, or purchasing a membership/plan.
   - **Open or Unrestricted Commercial Use** – Commercial use is allowed without restriction; the content is considered public or under an open license.
   - **Not Applicable** – The website does not specify any licensing or restrictions, or the service itself has no ruling over the content it hosts.
3. *Scraping Policy*
   - **No scraping and AI** – Explicitly prohibits scraping and AI for any content.
   - **No scraping** – Explicitly prohibits scraping, but no mention of AI.
   - **No AI** – Explicitly prohibits AI, but no mention of scraping.
   - **No scraping and AI conditionally** – Prohibits a part of the content from scraping and AI, or prohibits scraping and AI under certain conditions, such as the permission of robots.txt.
   - **Not Mentioned** – No explicit restrictions mentioned around scraping or AI in the Terms of Service.

D robots.txt Full Results

D.1 Summary Statistics

In the top 50 web domains from `small-en` and `medium-en`, we observe 3218 and 3879 agents, respectively. These observations cover 1,126,876 and 11,556,755 samples in `small-en` and `medium-en`, respectively.

D.2 Full Distributions

In table 6 and table 7, we see a very similar robots.txt analysis where the medium scale has about 10 times the observations as the total set scales up by 10 times. The dark gray background indicates that the "All Disallowed" rate, relative to the number of observations, is greater than or equal to 80%. We observe that the all AI-purposed robots have over 80% *All Disallowed* rates.

Table 6: Top results from robots.txt analysis for `small-en` scale's top 50 *base domains*, accounting for 96,436 attempted *full domains*, 81,273 successful robots.txt, and 1,126,876 samples. The full list of agents is not shown for conciseness. In this table, we only show agents with over 1,000 sample observations. The dark gray background highlights agents that have over 80% "All Disallowed" rate. For each agent, the number of observed cases is broken down by the number and percentage (relative to observed) of cases where all, some, or none were disallowed. "All Agents" row refers to an aggregation of all agents found in all the examined robots.txt. The aggregation rule is as follows: If for all agents, a robots.txt has *All Disallowed*, then the decision is *All Disallowed*. If for any agent in all agents, a robots.txt has *All Disallowed* or *Some Disallowed*, then a robots.txt has *Some Disallowed*. Otherwise, it has *None Disallowed*.

Agent	Observed	*All Disallowed*		*Some Disallowed*		*None Disallowed*	
		Count	% of observed	Count	% of observed	Count	% of observed
All Agents	1,126,876	6,442	0.6%	1,014,576	90.0%	105,858	9.4%
GPTBot	578,498	538,431	93.1%	40,028	6.9%	39	0.0%
*	475,139	18,595	3.9%	391,799	82.5%	64,745	13.6%
CCBot	353,324	313,920	88.8%	39,365	11.1%	39	0.0%
Bytespider	301,344	262,029	87.0%	39,274	13.0%	41	0.0%
googlebot-image	224,268	0	0.0%	224,166	100.0%	102	0.0%
claudebot	224,200	224,199	100.0%	1	0.0%	0	0.0%
Google-Extended	219,512	180,111	82.1%	39,367	17.9%	34	0.0%
SentiBot	219,365	180,086	82.1%	39,274	17.9%	5	0.0%
Baiduspider	204,497	35,762	17.5%	168,716	82.5%	19	0.0%
FacebookBot	183,430	144,102	78.6%	39,288	21.4%	40	0.0%
omgili	183,405	144,107	78.6%	39,274	21.4%	24	0.0%
Amazonbot	183,399	144,070	78.6%	39,297	21.4%	32	0.0%
omgilibot	183,118	143,820	78.5%	39,274	21.4%	24	0.0%
Googlebot-Image	180,355	32	0.0%	168,841	93.6%	11,482	6.4%
Bingbot	142,854	142,668	99.9%	40	0.0%	146	0.1%
Mediapartners-Google*	59,654	0	0.0%	0	0.0%	59,654	100.0%
GoogleContextual	59,231	0	0.0%	59,231	100.0%	0	0.0%
Twitterbot	52,649	6	0.0%	40,463	76.9%	12,180	23.1%
bingbot	49,452	7	0.0%	49,270	99.6%	175	0.4%
ClaudeBot	38,108	37,979	99.7%	91	0.2%	38	0.1%
Applebot-Extended	37,797	37,710	99.8%	55	0.1%	32	0.1%
PetalBot	36,696	36,647	99.9%	1	0.0%	48	0.1%
magpie-crawler	36,333	36,332	100.0%	0	0.0%	1	0.0%
applebot	36,269	0	0.0%	36,222	99.9%	47	0.1%
AdsBot-Google	28,599	25	0.1%	28,469	99.5%	105	0.4%
Yandex	15,974	401	2.5%	15,552	97.4%	21	0.1%
facebookexternalhit	15,678	7	0.0%	37	0.2%	15,634	99.7%
AdIdxBot	12,927	0	0.0%	12,905	99.8%	22	0.2%
Googlebot	12,303	26	0.2%	392	3.2%	11,885	96.6%
Pinterestbot	11,950	7	0.1%	11,891	99.5%	52	0.4%
ia_archiver	4,983	131	2.6%	4,695	94.2%	157	3.2%
anthropic-ai	1,739	1,689	97.1%	14	0.8%	36	2.1%
ImagesiftBot	1,636	1,592	97.3%	1	0.1%	43	2.6%
meta-externalagent	1,414	1,398	98.9%	2	0.1%	14	1.0%
PerplexityBot	1,409	1,223	86.8%	138	9.8%	48	3.4%
MJ12bot	1,033	982	95.1%	5	0.5%	46	4.5%

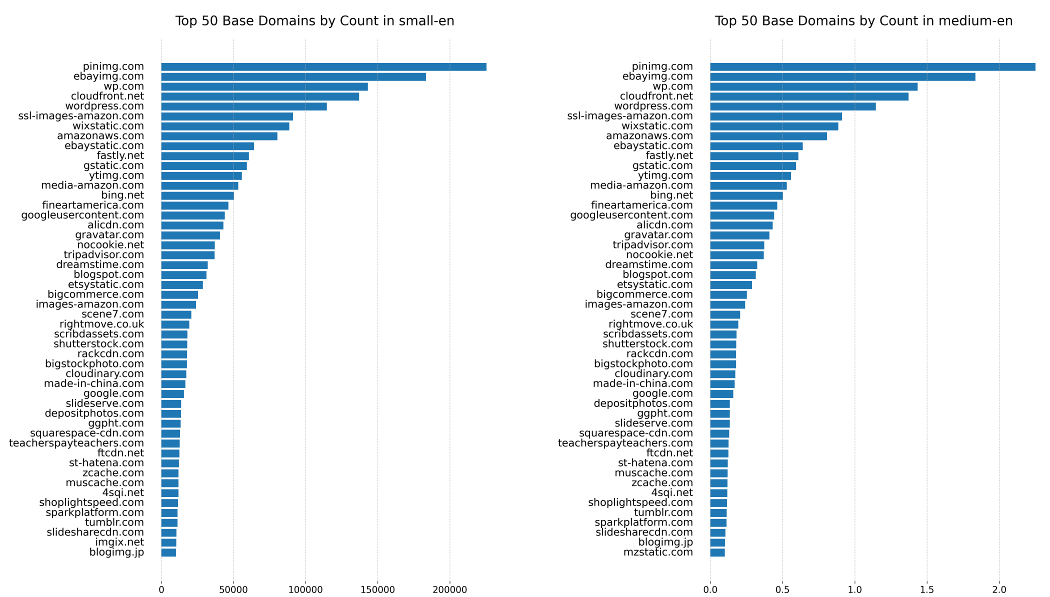
Table 7: Top results from robots.txt analysis for `medium-en` scale's top 50 *base domains*, accounting for 434,498 attempted *full domains*, 392,286 successful robots.txt, and 11,556,755 samples. The full list of agents is not shown for conciseness. In this table, we only show agents with over 10,000 sample observations. The dark gray background highlights agents that have over 80% "All Disallowed" rate. For each agent, the number of observed cases is broken down by the number and percentage (relative to observed) of cases where all, some, or none were disallowed. "All Agents" row refers to an aggregation of all agents found in all the examined robots.txt. The aggregation rule is as follows: If for all agents, a robots.txt has *All Disallowed*, then the decision is *All Disallowed*. If for any agent in all agents, a robots.txt has *All Disallowed* or *Some Disallowed*, then a robots.txt has *Some Disallowed*. Otherwise, it has *None Disallowed*.

Agent	Observed	*All Disallowed*		*Some Disallowed*		*None Disallowed*	
		Count	% of observed	Count	% of observed	Count	% of observed
All Agents	11,556,755	65,886	0.6%	10,521,922	91.0%	968,947	8.4%
GPTBot	5,781,111	5,378,225	93.0%	402,335	7.0%	551	0.0%
*	5,039,780	186,668	3.7%	4,296,202	85.2%	556,910	11.1%
CCBot	3,532,300	3,136,474	88.8%	395,388	11.2%	438	0.0%
Bytespider	3,014,323	2,619,564	86.9%	394,413	13.1%	346	0.0%
googlebot-image	2,239,424	0	0.0%	2,238,505	100.0%	919	0.0%
claudebot	2,238,757	2,238,756	100.0%	1	0.0%	0	0.0%
Google-Extended	2,203,460	1,807,713	82.0%	395,391	17.9%	356	0.0%
SentiBot	2,201,585	1,807,137	82.1%	394,404	17.9%	44	0.0%
Baiduspider	2,040,055	357,497	17.5%	1,682,293	82.5%	265	0.0%
Amazonbot	1,838,597	1,443,521	78.5%	394,671	21.5%	405	0.0%
FacebookBot	1,837,341	1,442,411	78.5%	394,581	21.5%	349	0.0%
omgili	1,836,966	1,442,343	78.5%	394,410	21.5%	213	0.0%
omgilibot	1,835,306	1,440,691	78.5%	394,406	21.5%	209	0.0%
Googlebot-Image	1,798,281	75	0.0%	1,683,359	93.6%	114,847	6.4%
Bingbot	1,431,194	1,429,300	99.9%	356	0.0%	1,538	0.1%
Mediapartners-Google*	597,643	1	0.0%	0	0.0%	597,642	100.0%
GoogleContextual	592,649	0	0.0%	592,649	100.0%	0	0.0%
Twitterbot	529,106	41	0.0%	407,550	77.0%	121,515	23.0%
bingbot	498,101	66	0.0%	496,491	99.7%	1,544	0.3%
ia_archiver	434,741	1,339	0.3%	431,807	99.3%	1,595	0.4%
ClaudeBot	384,024	382,664	99.6%	949	0.2%	411	0.1%
Applebot-Extended	380,818	380,218	99.8%	335	0.1%	265	0.1%
PetalBot	370,568	370,078	99.9%	18	0.0%	472	0.1%
magpie-crawler	366,942	366,927	100.0%	4	0.0%	11	0.0%
applebot	366,462	0	0.0%	365,972	99.9%	490	0.1%
AdsBot-Google	287,314	365	0.1%	285,986	99.5%	963	0.3%
Yandex	160,006	2,901	1.8%	156,800	98.0%	305	0.2%
facebookexternalhit	158,068	137	0.1%	310	0.2%	157,621	99.7%
AdIdxBot	129,314	1	0.0%	129,124	99.9%	189	0.1%
Googlebot	122,753	354	0.3%	3,677	3.0%	118,722	96.7%
Pinterestbot	119,439	112	0.1%	118,796	99.5%	531	0.4%
anthropic-ai	16,224	15,728	96.9%	196	1.2%	300	1.8%
ImagesiftBot	15,332	14,904	97.2%	18	0.1%	410	2.7%
meta-externalagent	14,945	14,790	99.0%	8	0.1%	147	1.0%
PerplexityBot	14,413	12,513	86.8%	1,435	10.0%	465	3.2%
MJ12bot	10,425	9,845	94.4%	41	0.4%	539	5.2%

Table 8: Annotation schema for each attribute considered for web-domain-level characteristics. CDN Provider refers to third-party providers of content delivery network (CDN) as a service, such as Amazon Web Services.

Attribute	**Types**
Category	Marketplace, CDN provider Blog, Website hosting Stock photo Content-sharing community
License Type	Personal/Noncommercial/Research Conditional commercial use Open/Unrestricted commercial use Not applicable
Scraping Policy	No scraping and AI conditionally No scraping and AI, No scraping Not mentioned

Figure 4: Distribution of the top 50 base domains in the `small-en` and `medium-en` splits of CommonPool. We observe the top 50 base domains only differ by one, where `small-en` has `imgix.net` and `medium-en` has `mzstatic.com`.



(a) Top 50 base domains in `small-en` by sample counts.

(b) Top 50 base domains in `medium-en` by sample counts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims and recommendations are made consistent with the approximated, and quantified results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss how our audit methods through Terms of Service can be limiting because of the lack of source url to the web pages that render the images. We also discuss the choice of two smaller scales of CommonPool because of the limited compute and storage space.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose all the search patterns, codebook of annotation, top web domains we annotate, the details on how we retrieve robots.txt, and watermark detection training details in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the code in the supplementary material. However, we do note that DataComp CommonPool is a url-text pairs dataset, which provides access to images through the urls. It's a known problem that some images can no longer be accessed or can change.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The only training part is finetuning MobileViTv2 for watermark detection, and we release all the details in both the training code and the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: In our only training experiment (for MobileViTv2), we do not include an error bar. All other off-the-shelf models are only evaluated, rather than trained. The main claim that watermark detection methods fall short for web-scraped images still holds even though MobileViTv2 experiments do not have the error bar. For annotating watermark samples in LAION-5B and CommonPool, we do include a confidence interval for the estimation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the compute resources for watermark detection in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: In this work, we do not release additional artifacts other than the annotations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We discuss the positive impact of our work on recommending more responsible image-text dataset releases. Since our work is a measurement and audit-focused effort, we do not release models or data that could potentially be misused.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work doesn't release dataset or models for downstream usage.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all the codebase and data we use from this work, including watermark detection datasets, methods, CommonPool's paper, and VLMs' paper.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We do document README.md in our codebase released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The human annotation done in this work is by the author team rather than crowdsourcing. Therefore, the item doesn't apply.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not Applicable

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is not used as a core method in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.